ARMY RESEARCH LABORATORY

An Evaluation of a Spoken Language Interface

by Paula P. Henry, Timothy J. Mermagen, and Tomasz R. Letowski

ARL-TR-3477   **April 2005**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

DESTRUCTION NOTICE—Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5425

An Evaluation of a Spoken Language Interface

Paula P. Henry, Timothy J. Mermagen, and Tomasz R. Letowski
Human Research & Engineering Directorate, ARL

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
April 2005	Final	

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
An Evaluation of a Spoken Language Interface	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
	62716AH70
Paula P. Henry, Timothy J. Mermagen, and Tomasz R. Letowski (all of ARL)	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
U.S. Army Research Laboratory Human Research & Engineering Directorate Aberdeen Proving Ground, MD 21005-5425	ARL-TR-3477

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBERS

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Speech recognition software has been proposed for inclusion in the Future Force Warrior system. DynaSpeak is a speech recognition software product developed by SRI International that has been shown to function well in quiet environments. However, the system has not yet been evaluated in noise environments representative of Army operational environments. This report summarizes a study that was conducted to evaluate the performance of the DynaSpeak software in the presence of two types of noise: steady state noise obtained from a tracked vehicle and impulse noise obtained from gunfire. Simultaneous recordings were made from 12 participants wearing two microphones: a Gentex noise-canceling boom microphone and a Temco HG-17 bone conduction microphone. The 12 participants verbalized call signs and commands which were analyzed by the speech recognition software. The results of the present study show that the DynaSpeak software functions well through the boom microphone in moderately high steady state noise (90 decibels A-weighted [dBA]) with error rates for words of 2% to 5%. However, the performance of the system with the same boom microphone in high steady state noise levels (110 dBA) or in moderately high impulse noise levels (90 dBA) does not meet the requirements for military applications. Error rates for words in these conditions ranged from 14% to 40%. If the DynaSpeak software is to be used in environments of impulse noise or high levels of steady state noise, the performance of the system must be enhanced.

15. SUBJECT TERMS
bone conduction; speech recognition

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Paula Henry
a. REPORT UNCLASSIFIED	b. ABSTRACT UNCLASSIFIED	c. THIS PAGE UNCLASSIFIED	SAR	30	19b. TELEPHONE NUMBER *(Include area code)* 410-278-5848

Contents

List of Figures

List of Tables

1. Introduction

The Future Force Warrior (FFW) program is considering the inclusion of a spoken language interface (also known as speech recognition software) for control of FFW system functions, including serving as the controller module for remote control of unmanned vehicles. One such product is the DynaSpeak[1] software developed by SRI[2] International. This system has been shown to function well in quiet environments but has not yet been evaluated in the presence of high noise and impulse noise environments representative of Army operational environments. The U.S. Army Research Laboratory (ARL) was approached by the Natick Soldier Center (NSC), Massachusetts, to conduct such an evaluation for a moving tracked vehicle (steady state noise) and gun fire (impulse noise) environments. This report summarizes the results of the evaluation.

The purpose of the study was to evaluate the effectiveness of the SRI DynaSpeak spoken language interface in high noise and impulse noise environments. This study sought to determine how well the SRI DynaSpeak software would recognize speech during a limited set of adverse operational conditions. It also sought to provide an answer as to whether the software performs more poorly, just as well as, or better than listeners with normal hearing in the same specified environments. This planned comparison was in case the DynaSpeak software did not perform well. In this case, a comparison to performance of listeners with normal hearing would provide information about whether the software or the test was the cause of the failure. The SRI DynaSpeak software, which meets the Army speech intelligibility criterion of 91% (Department of Defense, 1989), has been demonstrated to work well in quiet environments with various talkers. Through the use of noise-reduction algorithms, the software was expected to function well in steady state high intensity background noise generated inside and outside a moving vehicle. The function of the system in impulse noise environments was not projected to be acceptable for outgoing fire. However, such an acoustic environment may be prohibitively loud for direct speech communication. Therefore, the performance of the DynaSpeak software was compared with the recognition scores obtained from human listeners with normal hearing for data normalization.

2. Methods

Twelve adults (10 civilian and 2 military) with normal hearing sensitivity were recruited to serve as talkers and listeners in the study. Participants were between the ages of 18 and 51 years old

---

[1]DynaSpeak is a trademark of SRI International.

[2]Formerly Stanford Research Institute

(mean = 33). All participants had normal hearing defined as thresholds of 20 dB hearing level (HL) or better in both ears across 500 to 8 kHz (ANSI, 1996). The group of participants consisted of 10 men and 2 women.

The study had two portions: recording and listening. The recording portion was completed by all the participants before the listening portion was conducted. The recording portion of the study was conducted in the hostile environment simulator (HES) of building 518 at Aberdeen Proving Ground (APG), Maryland. The HES is a large (17.4 m long by 13.4 m wide and 6.7 m high) hall that has relatively low ambient noise levels measuring less than 30 dB A-weighted[3] (dBA). The measured reverberation time varies, depending slightly on the location of the microphone within the space, but is approximately 0.5 second in the frequency range from 500 Hz to 4000 Hz.

The listening portion took place in two sound-treated test booths in building 520, room 30, at APG. Two identical listening stations were set up with as many as two listeners at each station. This allowed for data collection from as many as four listeners at any given time.

At the beginning of the recording session, the participant was briefed about the research task and was asked to read and sign a volunteer agreement affidavit. The participant was then seated in the center of the HES and asked to verbalize lists of items. These lists included items from the Call Sign Acquisition Test (CAT) and simple commands used to operate unmanned vehicles (appendix A). The CAT consists of all 126 possible combinations of the 18 two-syllable Army phonetic alphabet and the seven one-syllable numbers from 1 through 8 (9 is not used because it is pronounced "niner" in radio communications); e.g. alpha 3, kilo 6 (Rao & Letowski, 2003; Rao, Letowski, & Blue, 2002). The list of commands was provided by SRI International and approved by representatives from MicroAnalysis & Design and NSC, who took part in the design of the study.

Participants were seated at a desk in the center of the HES, which contained two computer screens, a keyboard, and the two headsets to be used in the study. Before the participants were fitted with the headsets or were instructed in the use of the computer interfaces, they were presented with a list of the commands that would be used in the study. The participants were told that these were potential military-relevant commands that could be used with speech recognition software. They were asked to read aloud the list of 125 commands so that the experimenters could be assured that the participants were comfortable with how to say each of them and to answer any questions. The participants were also familiarized with the call signs. Variations in pronunciation were permitted for both the commands and the call signs (i.e., "route" in the command "route off" could sound like "rout" as in router or like "root"; Quebec in the call signs could be pronounced with a "qw" sound or a "k" sound at the beginning).

---

[3]A-weighted sound level measurements refer to measurements in which the sound levels of each frequency band are weighted in order to accommodate frequency-dependent changes in human auditory sensitivity at low intensity levels.

After the experimenter felt the participant was comfortable reading the call signs and the commands, the participant was fitted with hearing protection and the two headsets that would be used for recording. E•A•R[4] foam earplugs (noise reduction rating [NRR] = 30) and the combat vehicle crewmen (CVC) ear cups were used together to form an effective double hearing protection system.

Two microphones were used for making simultaneous recordings: a Temco HG-17 bone conduction microphone and a noise-canceling Gentex boom microphone extracted from a CVC helmet. The simultaneous use of both microphones was critical (identical verbalization of each phrase) for enabling meaningful comparisons of data obtained with each of the microphones. Figure 1 shows the two headsets that were used together in the present study. The HG-17 is a self-contained bone conduction communication system with a bone conduction microphone and bone conduction vibrators. This is a commercially available system not intended for acoustic environments. No modifications were made in the headset. The HG-17 bone conduction microphone was first placed on top of the talker's head. Participants were asked to adjust the headset so that the bone conduction microphone made good contact with their skulls. A helmet was not used in the study in order to ensure good contact between the bone microphone and the talker's head. After the bone microphone was placed on the talker's head, the position of the microphone and the static force pressing the microphone against the talker's head were adjusted to maximize the microphone output while ensuring the talker's comfort. If a helmet had been used, sufficiently good contact could not have been ensured for all the talkers. Such a bare head arrangement compromised low-frequency noise reduction by the bone microphone (through the absence of microphone shielding provided by a helmet). In the speech range (500 to 4000 Hz), natural reduction of external noise by the bone microphone is much higher than at low frequencies and the shielding action of a helmet is not as critical. Since the signal-processing algorithm in the DynaSpeak software is not affected by the upward spread of masking caused by low-frequency noise, the lack of noise shielding provided by the helmet should not affect the performance of the DynaSpeak software, but this lack of noise shielding could affect to some degree the live listener's performance. Therefore, it was determined that good contact between the bone microphone and the talker's head was more critical to the study than noise shielding of the bone microphone through use of a helmet.

The boom microphone in the CVC helmet includes wiring that is incorporated into the ear cups of the helmet. Since it was not desirable to use a helmet in the study, the boom microphone needed to be separated from the CVC helmet. The ear cups and boom microphone were extracted from the helmet and mounted into an earphone headband (see figure 1). This allowed the participant to wear a set of headphones with the boom microphone attached. Once the talker had the HG-17 headset in place, the boom microphone and headband were placed in such a manner as not to have the headband pressing directly on the bone conduction microphone. The use of the two systems together allowed for simultaneous recordings from both microphones.

---

[4]E•A•R is a registered trademark of Cabot Safety Corporation.

The typical placement of the headband from the earphones was just above the talker's forehead. The location of the boom microphone was adjusted so that the microphone tip was centered on and 0.25 inch from the talker's lips (Department of the Army, 1997).

Figure 1. Photographs of the two microphone systems used in the present study. (The left photograph is of the Temco HG-17 bone conduction system and the right photograph is of the boom microphone and ear cups from the CVC helmet mounted into a separate headband. No other headgear was used in the study.)

The Temco HG-17 communication system also includes a push-to-talk system (BM8). This device must be depressed to activate the bone conduction microphone. Once the microphone is activated, it stays on until the talker presses the button a second time or the talk circuit times out. The talkers were instructed in the use of the push-to-talk system and were instructed to reactivate the microphone after verbalizing approximately 25 items. The investigator sitting next to the talker confirmed that the microphone was always active during the recording periods.

The signal output levels from both microphones were set at the input to the computers to provide a maximum undistorted (no clipping) signal at the DynaSpeak input. The levels were set based on one of the experimenter's voices and maintained for all subsequent talkers. This procedure resulted in uniform undistorted recording levels for all the talkers and the bone microphone. In the case of the boom microphone, the recording level was set too high for the first two participants, resulting in excessive clipping. Therefore, these recordings were not used in the subsequent analysis of the data, and the input level for the boom microphone was slightly lowered for the rest of the talkers.

As stated before, the desk at which the talkers sat held monitors connected to two personal computers. These monitors displayed the DynaSpeak communication screens to the talker. The recordings were made from the two microphones simultaneously, one fed to each computer's sound card. This enabled two identical verbalizations to be analyzed through each of the two microphones. Two separate versions of the DynaSpeak software were used: one with a noise reduction algorithm designed for the bone microphone and one designed for the boom micro-

phone. Each microphone output was "hard wired" to the appropriate computer and the computers were labeled so that the participant knew which screen corresponded to which microphone.

The DynaSpeak software prompts the talker for each item (commands or call signs) that the talker is asked to record. The software also includes a volume unit meter that allows the talker to monitor the signal being received by the computer from the microphone. This proved especially useful in helping us determine whether the bone conduction microphone was transmitting a signal. Additionally, the software provided feedback to the talker as to what the DynaSpeak software determined the spoken phrase was, and it kept a running total of percent of words and sentences that were recognized correctly. The provision of feedback in speech recognition software allows the talker to repeat phrases or adjust his or her pronunciation to improve the system's ability to recognize the talker's voice. In the current evaluation, it was not desirable to have the talkers adjust their voices in any way. Therefore, the parts of the screen that displayed feedback and the running score were covered with paper to hide them from the talkers' view.

A wireless keyboard and two wireless receivers were used to synchronize the recordings on the two computers. The participant recorded a command or call sign by holding the control key (CTRL) on the keyboard and then vocalizing the phrase. When the talker released the control button and hit the "next" button (right arrow key), the software analyzed the recorded phrase and continued to the next item. The talker was instructed to speak with such effort as s/he normally would to communicate in the noise environment in which s/he was placed. The talker was also instructed to only repeat if s/he misspoke. For instance, if the command was "route off" and the talker said "route on," the item could be re-recorded in the same way it was initially before the next key was hit.

The recordings were analyzed by the DynaSpeak software on each system in real time as they were being recorded, and a cumulative percent recognition score for each system was maintained. The DynaSpeak software created the *.wav files with the recorded voice segment and a text file with the analysis of the performance of the verbalization. On rare occasion, the two computers were not synchronized. This did not occur very often, but when it did, the assisting experimenter stopped the participant and reversed the DynaSpeak software a number of items to re-synchronize the computers. At this point, the participant was asked to re-record the first item where the computers were noted by the experimenters or the participant to be out of synchronization, and the experiment was continued.

Two noises were selected for use in the evaluation: steady state noise recorded from a tracked military vehicle and military-relevant impulse noise (gunfire). They were representative examples of the steady state noise and the impulse noise present in military environments. The use of these two noises was requested by the FFW office, based on their military relevance and their representation of noises common to infantry Soldiers. The steady state noise was created from a digital recording of an M1A2 tank moving at approximately 10 mph on a flat gravel track. The noise was recorded through a microphone mounted outside the tank. The impulse noise was

a creation of multiple weapons fire relatively close to the listener. The weapons fire was taken from a battle scene in a movie with added simulated machine gun fire.

Two presentation levels were selected for each noise in the study: 90 and 110 dBA. These two noise levels represent lower and upper boundaries of predominant battlefield noises. The 90-dBA level is typical for acoustic environments in and around idling tracked vehicles. It is also a characteristic noise level for a person on the battlefield not being involved in direct action (receiving and returning direct fire). The level of 110 dBA represents the level common inside and in proximity to rapidly moving tracked vehicles. It is also a reasonable estimate of the average noise exposure level of a person involved in an exchange of direct fire. In other words, the levels of 90 and 110 dBA represent moderate and high level noise exposure on the battlefield, respectively.

Participants completed a total of 16 recordings (2 microphones x 2 noise conditions [impulse noise, steady state tank noise] x 2 noise levels [90 and 110 dBA] x 2 sets of stimuli [call signs, commands]). The ordering of the recordings was based on a Latin square design where three participants followed any given order. The ordering of items within a block was randomized before the study began, and the same randomization was used for each of the participants. Both microphone recordings were obtained simultaneously for each talker, resulting in eight total blocks of recordings per talker. The two sets of stimuli (call signs and commands) were presented in separate blocks to allow for analysis to be conducted on each set of stimuli separately. Both recordings were played back in the test environment (HES) at two different levels: 90 and 110 dBA. For the impulse noise recording, the presentation level was set, based on the machine gun noise peaks in the noise file. The participants were required to wear E•A•R foam plugs in addition to the ear cups from the CVC helmet in order to ensure adequate protection for hearing. The foam plugs were inserted into the participants' ears by the participant with experimenter supervision or by the experimenter, depending on participant's preference as well as the experimenter's determination of a proper fit.

Because the HG-17 communication system and the ear cups were used simultaneously, an adequate seal between the ear cups and the listener's head was not obtained. Therefore, we elected to use a double hearing protection scheme with the E•A•R foam plugs serving as primary hearing protectors and providing approximately 15 to 30 dB of hearing protection and the ear cups serving as secondary hearing protectors.

During recordings, the participants were provided several breaks in a quiet environment to rest their voice and hearing. Calculations of total noise exposure were made to ensure a safe recording environment for the participants. All participants completed the recording sessions before any completion of the listening sessions.

Once the recordings were complete, the results of the DynaSpeak software were evaluated. The evaluation focused on the number of words and sentences that the DynaSpeak software did not correctly recognize. Separate values were obtained for the commands and call signs. The

conditions in which the recordings were made were documented and sound files were organized for subsequent playback to the participants who served as human listeners. The selection of the recordings to be used for the listening portion of the study represented the worst case scenario. First, the talker whose recordings resulted in the greatest number of errors by the DynaSpeak software was chosen. Next, three conditions for the boom microphone were chosen for presentation to the listeners. These three conditions represented recordings in which the software performed poorly. Knowing that the human listeners should outperform the DynaSpeak software, this was the best scenario for comparing relative effects of the test conditions.

For the human listener evaluation, the recordings obtained from the talker in any given condition were combined into a single sound file. We accomplished this by taking all the *.wav files and adding a sample of the background noise from that recording between items. We obtained the samples of background noise by making noise-only recordings through each of the microphones in each of the noise conditions. The length of the background noise sample placed between the items varied, based on the type of stimuli. For the CAT items, the background noise sample lasted 2 seconds. For the commands, the background noise sample lasted 10 seconds. The ordering of the items was randomized via a computer program designed for this purpose. The same randomization was used for all participants.

The listening portion of the study was performed in ARL's sound-treated booths in room 30 in building 520 at APG. Two separate listening stations were established. Personal computers were used to play the sound files through SoundForge[5] software. The output of the sound cards was first routed to a Rane HC6 headphone amplifier and then to two pairs of AKG (Acoustics GmbH) K240 DF[6] studio monitor headphones. Individual volume controls were provided for each listener so that s/he could adjust the output to the headphones to a comfortable listening level. Each participant listened to recordings made in three conditions: the CAT recorded in 110 dBA of steady state tank noise, the CAT recorded in 90 dBA of impulse noise, and the commands recorded in 110 dBA of tank noise. The participants were told what they were going to hear and how they were going to record their responses.

If the recording was of the CAT call signs, the listeners were instructed to identify each phonetic alphabet letter and number combination (call sign) by writing the letter followed by the number. For example, if s/he heard "alpha six," s/he should record "A6." If the recording was of the command phrases, the listeners were instructed to write the whole phrase. The listener was encouraged but not required to abbreviate some of the words to save time. A new record sheet was used for each condition, and each record was coded with condition and participant numbers. The ordering of the recordings played to the listeners followed a Latin square design where every four listeners followed a different order.

Twelve participants completed the listening portion of the study. All the listeners completed the recording portion of the study before they completed the listening portion. The record sheets were scored on the basis of percent of items (call signs or commands) recognized correctly. The

---

[5]SoundForge is a trademark of Sonic Foundry.

[6]not an acronym

performance of the participants in the listening conditions was then compared to that of the DynaSpeak software.

3.   Results and Discussion

3.1   Results From the DynaSpeak System

The items read by each talker were analyzed by the DynaSpeak system in real time as the recordings were being made. After recordings of any given talker were made, the cumulative percentage of errors was extracted from the data log files contained in each talker's folder for each of the microphone, stimulus set, noise, and intensity level combinations. Word level and sentence (phrase) level errors were calculated. As per the recommendation from the FFW office, presented discussion of the data is focused on the percentage of words that were not recognized correctly (percent of errors).

The percentage of word errors for each of the conditions after being analyzed by the DynaSpeak software is shown in figures 2 and 3. Figure 2 shows the results for the boom microphone, and figure 3 shows the results for the bone conduction microphone. Three general and consistent patterns can be seen in these figures. First, the results demonstrate that the software functioned better at the lower noise level (90 dBA) than at the higher noise level (110 dBA). This is the expected result, but the important characteristic is the size of the difference.
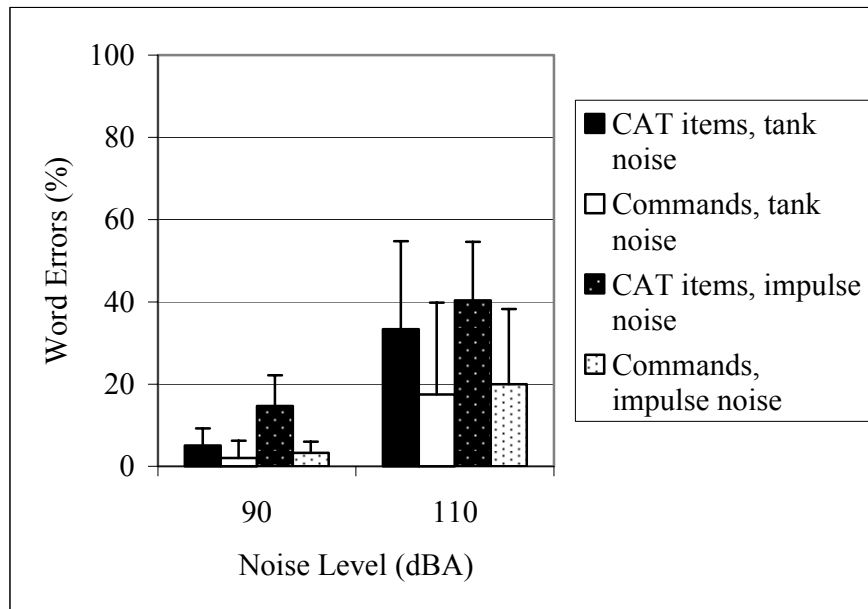

Figure 2. Results from the DynaSpeak software on the recordings made through the boom microphone, averaged across listeners and showing the percentage of errors of words from each condition. (Error bars indicate +1 standard deviation.)
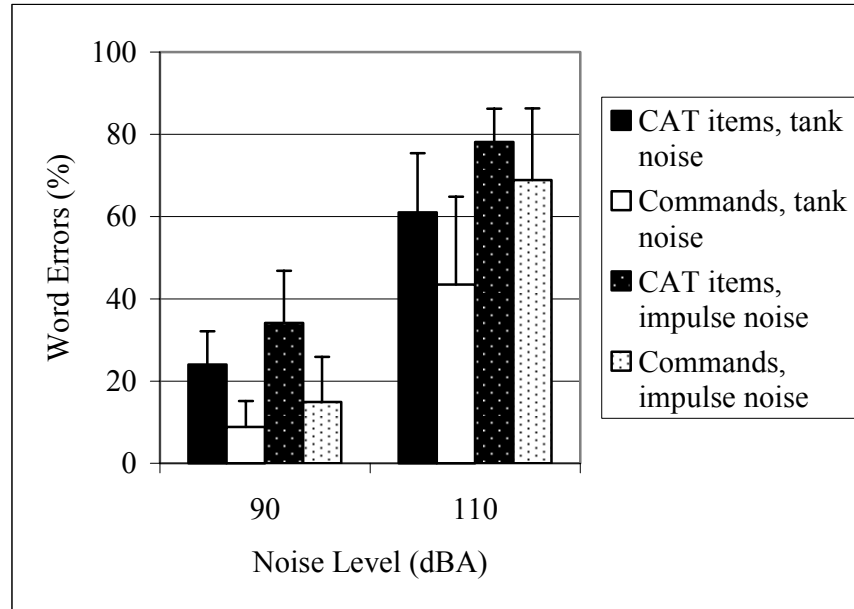
8

Figure 3. Results from the DynaSpeak software on the recordings made through the bone conduction microphone, averaged across listeners and showing the percentage of errors of words from each condition. (The error bars indicate +1 standard deviation.)

Second, small but consistent differences appear between the two types of noises (tank and impulse). The software made fewer errors in speech recognition with the relatively steady state tank background noise than with the impulse noise. This is probably attributable to the noise reduction algorithm's ability to sample and more effectively cancel a noise that is steady state in nature. It is quite possible that further fine tuning of the processing algorithm may result in indifference to the type of noise.

Third, the software functioned better for the stimuli that consisted of commands rather than call signs. This is probably because of the presence of contextual effects in commands that are not present in call signs. The software uses a limited set vocabulary and is testing for the presence of specific words and word combinations. The probability of correct guessing of the CAT phrase is less than 1%, assuming that the DynaSpeak system expects only a CAT phrase and much less than 1% if not. Conversely, in the commands set, there are some unique phrases, in which after the leading word is recognized, there is only one option for the next word.

The difference in performance between the two noise levels, 90 and 110 dBA, is most likely attributable to differences in signal-to-noise ratio (SNR) at the input to the microphone. Recall that the talkers were instructed to speak as they would in real life in order to communicate in these specific noise environments. It is natural for a talker to raise his or her voice in the presence of noise in order for his or her speech to be understood by a listener. The talker would raise his or her voice in both noise conditions; however, the resultant SNR may differ. The projected lower SNR provided to the system in the 110-dBA condition would explain to some

9

degree the difference in error rates between the two noise levels. SNRs cannot be accurately measured directly because the speech levels are less than the levels of the background noise, thus resulting in a negative SNR. For several of the talkers, an indirect measurement was made of their vocal effort in the presence of 110 dBA noise. We measured this by playing the noise and then turning it off abruptly when the talker was speaking. The intensity levels of these talkers at the location of the boom microphone were between 85 and 95 dBA.

Please note that talkers are unable to maintain a normal speech level (vocal effort) in noise even if asked to do so. This is called the "Lombard effect". The voice level unconditionally increases with the increase in noise level, especially for noise levels exceeding 70 dBA. This increase is accompanied by a slower rate of speech, improved pronunciation, and changes in the spectral envelope of speech. All these changes result in Lombard speech being much easier to understand than normal speech by the listeners with normal hearing and by automatic speech recognition systems (Chen, 1988; Chi & Oh, 1996; Junqua, 1989; Letowski, Frank & Caravella, 1993; Summer, Pisoni, Bernacki, Pedlow, & Stokes, 1988). This improvement in speech recognition can be as large as 15% for a noise level of 90 dBA and an SNR = 10 dB (Chi & Oh, 1996). Therefore, the increase in noise level accompanied by Lombard speech may result in poorer SNR but not necessarily poorer speech intelligibility. This is obviously not true for high noise levels in which the level of Lombard speech has reached its limits because of natural limitations of the human vocal system. It is believed that the critical noise level is about 90 to 100 dBA, depending on the type of noise.

The differences seen between the two noise conditions (tank and impulse) are most likely attributable to the difference in the ability of the noise-reduction algorithms to eliminate noises of different types. For a relatively steady state noise, a sample of the noise can be taken and an algorithm created to minimize its impact on the speech signal. The noise recorded from a tank does not change significantly over time. Therefore, at any given moment, the background noise is relatively the same and remains so across different speech items. This is not necessarily true with impulse noise. In this case, the noise varies in time and is less predictable in nature. Therefore, an algorithm to reduce the impact of impulse noise needs to sample the background noise much more frequently and does not necessarily improve its performance in time because of constant changes in the nature of the noise. In addition, the presence of a speech-like background mixed into the impulse noise could to some degree confuse the DynaSpeak recognizer. It is unlikely that this was a major factor in the present study because the speech levels were 30 or more decibels below the level of the impulse noise. However, the human voice and speech "babble" need to be used as interfering noises in future studies to ensure the DynaSpeak's effectiveness in such environments.

The differences seen between the two sets of stimuli (call signs and commands) are probably attributable to their differences in contextual cues. Speech recognition systems often operate by comparing individual words against a list of potential phrases containing that particular word. The number of words that the software can recognize is often referred to as the software's

vocabulary, and the rules that bind individual words in phrases (choices) as the software's grammar. The size of the vocabulary for any particular group of phrases for the DynaSpeak system is not known, but it is reasonable to assume that the probability of occurrence of a specific combination of words is smaller for commands than for call signs. In the set of commands provided for the study, some words appeared in only a small number of phrases. For example, the word "cad" only correlated to one phrase, "cad_r_g". In this case, the number of potential phrases containing the word "cad" was one. So, as long as the word "cad" was recognized correctly, the entire phrase should also be recognized correctly. On the other hand, for some words, the number of possible phrases containing that particular word could be quite large. For instance, the word "change" was the first word in 12 different phrases, so the recognition of this particular word did not ensure the software's ability to recognize the entire command correctly. Therefore, the chance that the software will recognize a complete command correctly varies by individual command set.

For the call signs, any given alpha item could be paired with any given number, and the chances of the software recognizing the entire call sign correctly (both alpha and numeric) were the same for each call sign. Recall that there were 18 alpha and 7 numeric items in the set that the talker read, and they could be presented in any alpha-number combination. Therefore, the CAT tested the performance of the DynaSpeak system, based on an equal and low probability of each item. The probability was further decreased because the DynaSpeak vocabulary for call signs included all 26 alpha and 10 numeric items. The call sign test condition was probably very similar to that created by the set of command phrases using the word "change" as the initial word. However, overall, the choices the DynaSpeak system had to make for the command set were based on fewer possibilities than those produced by the CAT items because of unequal and sometimes high probability of an item's occurrence. Therefore, any change in the number and type of phrases used in the command set would likely affect performance of the system even if all the individual words were in the DynaSpeak vocabulary. Conversely, replacing specific items in the CAT set should not affect the DynaSpeak performance as long as these items are in the DynaSpeak vocabulary. Thus, the score for the CAT may be considered as an indicator of DynaSpeak's average ability to recognize individual words from its vocabulary, whereas the score for the command set may be considered as an indicator of how the pre-selected set of items maximizes current DynaSpeak capabilities.

The use of two microphones allowed for an evaluation of how the software would perform with differences in input SNR, noise-canceling options, and differences in available frequency bandwidth of the talker's voice. A comparison of figures 2 and 3 reveals that the boom microphone input was processed better by the DynaSpeak system than by the bone microphone input. This can be attributable to several factors: (1) the bone microphone program was not processing signals equally well as the boom microphone program, (2) the bone microphone provided a much poorer signal than the boom microphone and this difference could not be compensated by software adjustments, (3) a lack of noise shielding for the bone microphone

affected its performance, (4) the sound cards in both computers differed in their performance, and (5) the placements of the boom and bone microphones differed in their optimization. All the factors listed could simultaneously affect the data; however, the greatest contributor seems to be item (2). Noise-canceling microphones are mature technology and the Gentex noise-canceling microphone used in this study is recognized as one of the best on the market. Conversely, the bone microphone technology is just entering the market and cannot be described as a mature technology. The Temco microphone was the only reasonable choice several months ago, but it has growing competition. Therefore, the presented data can only be considered as the data describing performance of two specific transducers but cannot be generalized to all "boom microphones" and "bone microphones" as categories of transducers.

As stated earlier, a decision had to be made as to which aspect of experiment instrumentation was more critical to the current study: (1) maintaining good contact between the bone microphone and the talker's skull or (2) shielding the bone microphone through use of a helmet. Although the negative effect of lack of shielding on the bone microphone data cannot be totally discarded, the data presented in figures 1 and 2 show clearly that the repeatable type of contact between the bone microphone and the talker's skull has been ensured. Data variabilities obtained for the boom and bone microphones during various test conditions are very similar, thus indicating similar repeatability. The interfacing of both types of microphones can therefore be considered as equally reliable.

3.2   Results From Listeners With Normal Hearing

Individuals with normal hearing sensitivity were recruited to listen to several of the recordings and provide speech recognition data. This additional task was conducted in order to compare the software's performance with what would be expected by humans in the same environment. The results of an evaluation of speech recognition software without a comparison to what would be expected from listeners with normal hearing operating in the same acoustic environment could not be labeled as good, fair, or poor since there would be no basis of comparison.

It is well known that individuals with normal hearing can recognize speech at a negative SNR, which indicates that the intensity of the speech is less than that of the noise. Furthermore, speech recognition systems typically require a positive SNR for optimal performance. However, given the use of noise-reduction algorithms, noise-canceling microphones, and two microphones with different spectral content, it is unknown to what extent the performance of the software would agree with that of listeners with normal hearing.

The percentage of errors obtained from the DynaSpeak software for each individual talker is presented in appendix B. Three boom microphone recordings of a single voice presenting the worst case scenario were selected for human subject evaluation. These recordings were the CAT recorded in 110 dBA of tank noise, the CAT recorded in 90 dBA of impulse noise, and the commands recorded in 110 dBA of tank noise. The reason for selecting the worst case scenario for human subject listening was twofold: (1) there was relatively good overall performance of

the DynaSpeak system with the boom microphone input with a 90-dBA noise level and (2) there was a need to determine how poor the DynaSpeak performance can be in comparison to human hearing performance when there is a mismatch between the DynaSpeak software and human voice. Please note that if conditions when the DynaSpeak software performed quite well were selected, any higher performance by listeners with normal hearing would not be of much interest. Figure 4 shows the average results from the listening tasks, along with the performance of the DynaSpeak system for the same three recordings for comparison.

As seen in figure 4, the listeners performed far better than the DynaSpeak software with the talker whose voice resulted in the poorest performance of the software, as shown by lower percentages of errors. As stated previously, the fact that listeners with normal hearing generally function better than a speech recognition system is not in itself surprising as this has been demonstrated previously. What is of interest is the degree to which the two differed in this particular evaluation. For the three recordings that were chosen, the difference in performance was on the order of 30% to 40%. This shows that the software was substantially worse than what would be expected from a listener with normal hearing. However, the performance of the listeners with normal hearing was not acceptable, which indicated that although the system's performance should be improved, one cannot expect it to perform better than an error rate of 30% to 60% in 110 dBA noise (sentences) for some type of voices.
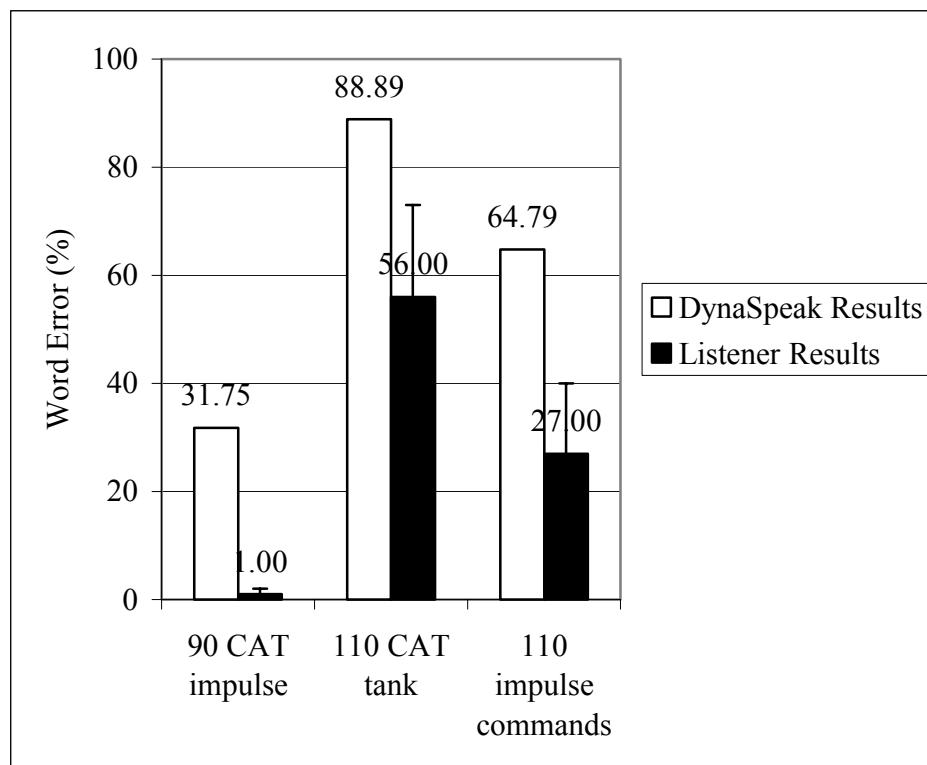

Figure 4. Average results for the 12 participants who completed the listening task compared to the results from the DynaSpeak software for the talker who resulted in the software's worst performance. (All results are from recordings made through the boom microphone. Error bars indicate +1 standard deviation.)

4. Conclusions

The DynaSpeak system by SRI International has been demonstrated previously by the developers to function well in quiet environments. The results of the present study show that the system also functions commendably well in moderately high steady state noise levels (90 dBA). However, the performance of the system in high noise levels (110 dBA) or in moderately high impulse noise levels (90 dBA) does not meet the requirements for military applications. If the DynaSpeak software is to be used in these conditions, the performance of the system must be enhanced. Otherwise, the software application should be limited to no more than 90 dBA of steady state noise. Finally, the software needs to be tested in a background noise consisting of human voices to determine its robustness in this type of environment.

5. References

American National Standards Institute. *Specification for Audio Meters*; ANSI 3.6-1996; New York, 1996.

Army Hearing Conservation Policy. Amendment to DA PAM 40-501, Washington, D.C., 1994.

Chen, Y. Cepstral domain talker stress compensation for robust speech recognition. *IEEE Trans. ASSP* **1988***, 36,* 433-439.

Chi, S.; Oh, Y. Lombard effect compensation and noise suppression for noisy Lombard speech recognition. *Proceedings of the Fourth International Conference on Spoken Language Processing (CDROM, vol. 4, Paper #065),* Philadelphia, PA, 3-6 October 1996.

Department of the Army. Technical Manual: Operator's Manual: Intercommunication Set, Vehicular AN/VIC-3(V), TM 11-5830-263-10, Washington, D.C., 1997.

Department of Defense. Military Standard: Human engineering design criteria for military systems, equipment and facilities. MIL-STD-1472D, Washington, D.C., 1989.

Junqua, J. C. The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America* **1989***, 85* (2)*,* 849-900.

Letowski, T.; Frank, T.; Caravella, J. Acoustical properties of speech produced in noise presented through supra-aural earphones. *Ear and Hearing* **1993***, 14,* 332-338.

Rao, M. D.; Letowski, T. R. Speech intelligibility of the Call Sign Acquisition Test (CAT) for Army communication systems. *Audio Engineering Society Convention Paper 5835.* Presented at the 114th Convention 2003, May 22-25, Amsterdam, The Netherlands, 2003.

Rao, M. D.; Letowski, T. R.; Blue, M. Subjective-objective assessment of speech intelligibility for the Callsign Acquisition Test (CAT). Unpublished data analysis, 2002.

Summer, W. V.; Pisoni, D. B.; Bernacki, R. H.; Pedlow, R. I.; Stokes, M. A. Effects of noise on speech production: acoustic and perceptual analyses. *Journal of the Acoustical Society of America* **1988***, 93* (1)*,* 917-928.

INTENTIONALLY LEFT BLANK

Appendix A. Commands and Call Signs Used as Speech Stimuli

A1. Commands:

change one down to c_two reporting display
v_map one
observation point
linkup point
action taken observing
change two up to tactical situation display
points off
change one up to situational awareness sensor
display
zoom fifty percent
shelter surface
done
mortar wheeled
way point
action taken request resupply
vehicle up on
zoom to oh point twenty five
zoom to one
air cavalry
change three down to rsta display
enemy activity attacking
signal
battle position section
change one up to system status display
squad
communications
release point
friendly neutral on
objective area
infantry
motorized infantry
battalion
cancel
coordination point
regiment
named area of interest
zoom to three point oh
zoom to two point zero
wheeled mortar
passage point
enemy activity engaging
cancel

o_n_c
spot report
change three down to target acquisition display
platoon
action withdrawing
enemy activity request resupply
engagement area
change one down to w_c_a alert list display
route off
undo last command
cad_r_g
annotation on
enemy activity river crossing
action taken reconnaissance
mortar
aircraft generic
edit type
way point
add to target queue
unit
team
areas on
infantry motorized
declutter
movement one hundred degrees
done
underground shelter
friendly off
map overlap on
generic aircraft
engineer
down
battle position platoon
targeted area
enemy unknown off
zoom two hundred percent
change three down to c_two reporting display
n_b_c smoke
change one up to s_a sensor display
friendly
zoom to zero point two five
n_b_c unknown
enemy

17

map on
done
enemy on
change two down to unmanned asset control
display
cancel
zoom three hundred percent
company
unknown
medical
cursor center off
action acquire
view
zoom out
done
change two up to unmanned asset control
display
enemy nationality friendly
mechanized infantry
maintenance
v_p_f
check point
headquarters
n_b_c area biological
change one down to s_a sensor display
recon
zoom to three point zero
target reference point
field artillery
reconnaissance
passage point
pan right
start point
enemy activity guarding
change three up to r_s_t_a display
enemy activity defending
critical point
load
general point
pan left
done
zoom to one point five
n_b_c area unspecified

A2. Call signs:

quebec four
oscar two
hotel three
whiskey three
bravo eight
hotel five
hotel four
whiskey four
x-ray six
charlie four
lima four
yankee three
victor five
victor six
papa two
tango two
quebec five
quebec eight
yankee five
bravo five
foxtrot two
whiskey two
foxtrot five
Kilo eight
hotel one
hotel six
tango three
foxtrot one
lima two
papa four
x-ray eight
charlie one
bravo two
echo four
quebec six
bravo six
foxtrot six
foxtrot eight
oscar three
kilo two
lima three
delta one
oscar four
delta four
alpha four
tango one

whiskey six
foxtrot four
quebec three
victor two
tango four
yankee eight
x-ray three
x-ray five
oscar six
alpha two
lima six
tango six
papa three
bravo one
x-ray four
bravo four
whiskey five
echo two
echo five
quebec one
echo three
charlie six
foxtrot three
quebec two
victor four
hotel two
yankee two
charlie five
kilo four
yankee four
alpha six
delta five
whiskey one
papa one
tango five
lima five
charlie two
echo eight
oscar five
echo one
yankee one
bravo three
kilo three
zulu five
alpha five
zulu two

delta three
delta six
victor three
papa six
kilo five
yankee six
delta two
charlie three
alpha three
charlie eight
hotel eight
echo six
zulu six
kilo six
tango eight
x-ray two
alpha one
zulu eight
lima eight
oscar eight
lima one
delta eight
whiskey eight
oscar one
zulu three
papa five
victor eight
zulu four
papa eight
x-ray one
alpha eight
victor one
zulu one
kilo one

19

INTENTIONALLY LEFT BLANK

Appendix B. Results From the DynaSpeak Software for Individual Talkers

Table B-1. Results from the DynaSpeak software for individual talkers. (Recordings were made from the Gentex noise-canceling boom microphone.)

				Error for Words				
subj no.	CAT_90_ tank	CAT_110_ tank	commands_ 90_tank	commands_ 110_tank	CAT_90_ impulse	CAT_110_ impulse	commands_ 90_impulse	commands _110_impulse
1B	3.17	12.70	0.59	5.03	6.35	33.33	2.07	10.65
2B	7.14	39.29	0.30	19.53	17.06	35.71	2.37	7.40
3	4.37	27.38	1.18	13.02	10.32	28.17	2.96	8.58
4	5.16	50.00	0.30	19.23	12.30	57.14	1.18	31.95
5	0.40	14.29	0.30	8.28	9.52	43.65	3.55	34.62
6	7.94	36.11	2.66	18.05	21.43	52.38	6.80	36.09
7	1.98	13.10	0.59	2.96	4.76	23.02	1.18	6.80
8	2.78	21.83	0.30	2.37	7.54	32.14	1.78	5.33
9	0.40	32.54	0.89	12.13	15.87	28.97	0.89	10.95
10	3.17	20.63	0.59	4.44	13.89	37.30	2.66	6.51
11	9.13	43.25	1.78	20.41	17.46	39.68	4.14	15.98
12	15.08	88.89	15.08	84.91	31.75	73.02	10.06	64.79
avg	5.06	33.33	2.05	17.53	14.02	40.38	3.30	19.97
sd	4.21	21.39	4.17	22.28	7.49	14.24	2.68	18.26

Table B-2. Results from the DynaSpeak software for individual talkers. (Recordings were made from the Temco HG-17 bone conduction microphone.)

				Error for Words				
subj no.	CAT_90_ tank	CAT_110_ tank	commands_ 90_tank	commands_ 110_tank	CAT_90_ impulse	CAT_110_ impulse	commands_ 90_impulse	commands_ 110_impulse
1B	45.24	78.57	17.75	43.82	51.57	82.82	18.79	66.86
2B	15.87	41.27	6.12	23.16	32.81	64.29	7.10	44.89
3	16.27	73.02	5.62	58.28	23.41	83.73	8.88	88.17
4	24.61	47.22	4.64	29.91	31.50	78.17	7.69	68.93
5	20.63	67.46	12.43	58.28	36.11	88.10	25.44	90.83
6	32.54	67.46	22.49	63.02	62.30	80.95	39.35	77.51
7	25.00	60.32	7.10	26.33	17.86	76.19	5.92	69.23
8	21.83	46.54	2.37	17.75	30.16	66.80	6.96	46.27
9	20.08	61.54	4.58	44.08	30.71	75.79	7.60	62.13
10	25.78	61.11	10.20	47.63	44.14	79.76	27.76	70.43
11	26.89	55.04	12.78	30.77	27.78	69.47	11.47	52.30
12	29.39	86.61	13.02	91.42	46.83	89.29	27.57	91.88
avg	25.34	62.18	9.93	44.54	36.27	77.95	16.21	69.12
sd	7.99	13.49	6.01	20.97	12.67	7.93	11.23	16.15

22

1   ARMY RSCH LABORATORY - HRED
     ATTN AMSRD ARL HR MJK MS D BARNETTE
     JFCOM JOINT EXPERIMENTATION J9
     JOINT FUTURES LAB
     115 LAKEVIEW PARKWAY SUITE B
     SUFFOLK VA 23435

1   ARMY RSCH LABORATORY - HRED
     ATTN AMSRD ARL HR MQ M R FLETCHER
     US ARMY SBCCOM NATICK SOLDIER CTR
     AMSRD NSC SS E   BLDG 3 RM 341
     NATICK MA 01760-5020

1   ARMY RSCH LABORATORY - HRED
     ATTN AMSRD ARL HR MT DR J CHEN
     12350 RESEARCH PARKWAY
     ORLANDO FL 32826-3276

1   ARMY RSCH LABORATORY - HRED
     ATTN AMSRD ARL HR MS MR C MANASCO
     SIGNAL TOWERS   RM 303A
     FORT GORDON GA 30905-5233

1   ARMY RSCH LABORATORY - HRED
     ATTN AMSRD ARL HR MU M SINGAPORE
     6501 E 11 MILE RD MAIL STOP 284
     BLDG 200A 2ND FL RM 2104
     WARREN MI 48397-5000

1   ARMY RSCH LABORATORY - HRED
     ATTN AMSRD ARL HR MF MR C HERNANDEZ
     BLDG 3040 RM 220
     FORT SILL OK 73503-5600

1   ARMY RSCH LABORATORY - HRED
     ATTN AMSRD ARL HR MW E REDDEN
     BLDG 4 ROOM 332
     FT BENNING GA 31905-5400

1   ARMY RSCH LABORATORY - HRED
     ATTN AMSRD ARL HR MN R SPENCER
     DCSFDI HF
     HQ USASOC BLDG E2929
     FORT BRAGG NC 28310-5000

1   DR THOMAS M COOK
     ARL-HRED LIAISON
     PHYSICAL SCIENCES LAB
     PO BOX 30002
     LAS CRUCES NM 88003-8002

1   US ARMY SAFETY CTR
     ATTN CSSC SE
     FORT RUCKER   AL 36362

ABERDEEN PROVING GROUND

1   DIRECTOR
     US ARMY RSCH LABORATORY
     ATTN AMSRD ARL CI OK (TECH LIB)
     BLDG 4600

1   US ATEC
     RYAN BLDG
     APG-AA

1   DIRECTOR
     US ARMY RSCH LABORATORY
     ATTN AMSRD ARL CI OK TP S FOPPIANO
     BLDG 459

1   DIRECTOR
     US ARMY RSCH LABORATORY
     ATTN AMSRD ARL HR MB   J NAWLEY
     BLDG 459

1   DIRECTOR
     US ARMY RSCH LABORATORY
     ATTN AMSRD ARL HR M F PARAGALLO
     BLDG 459